MCMC 2: Lecture 3
SIR models (2)

Phil O'Neill   Theo Kypraios

School of Mathematical Sciences

University of Nottingham

Contents

- 1. Non-Markov SIR epidemic model
- 2. Debugging tips
- 3. What to do with MCMC output
- 4. What can be estimated?

Contents

- 1. <u>Non-Markov SIR epidemic model</u>
- 2. Debugging tips
- 3. What to do with MCMC output
- 4. What can be estimated?

1.Non-Markov epidemic model

Population of N individuals

At time t there are:

S(t) susceptibles

I(t) infectives

R(t) recovered / immune individuals

Thus S(t)+I(t)+R(t) = N   for all t.

Initially (S(0),I(0),R(0)) = (N-1,1,0).

1.Non-Markov epidemic model

- Each infectious individual remains so for a length of time T_I drawn from $f(x \mid \theta)$. Note that θ might have several components.

- During this time, infectious contacts occur with each susceptible according to a Poisson process of rate β / N.

- Thus overall infection rate is $\beta S(t)I(t) / N$.

- Model parameters are β and θ .

1.Non-Markov epidemic model

- Popular choices for the infectious period distribution T_I include the Gamma distribution and the Weibull distribution.

- If T_I is the Exponential distribution then we have the general epidemic = Markov SIR epidemic.

- For non-Exponential T_I then $\{(S(t),I(t)): t \geq 0\}$ is not a Markov process.

1. Non-Markov epidemic model

- Suppose we observe n removals at times

$$r_1 \leq r_2 \leq r_3 \leq \ldots \leq r_n.$$

- The inference problem is to estimate the model parameters β and θ.

- In a Bayesian setting, this means we need to find (or sample from) the posterior density

$$\pi(\beta, \theta \mid r_1, r_2, \ldots, r_n)$$

1. Non-Markov epidemic model

- As for the general epidemic, the likelihood

$$\pi(r_1, r_2, ..., r_n \mid \beta, \theta)$$

is very hard to compute.

- Again our solution is to introduce infection times as extra variables to give a tractable augmented likelihood.

1. Non-Markov epidemic model

Augmented Likelihood

- Let b be the label of the last removal time, i.e. $r_b \geq r_k$ for all k = 1, ..., n.

- Given removal data, b is observed and fixed for any given labelling.

- Define a as the label of the first infection time, i.e. $i_a < i_k$ for all k ≠ a.

- Given removal data, a is unknown.

1. Non-Markov epidemic model

- Define $\mathbf{r} = (r_1, r_2, \dots, r_n)$
- Define $\mathbf{i} = (i_1, i_2, \dots, i_{a-1}, i_{a+1}, \dots, i_n)$
- Let $f(x \mid \theta)$ denote the probability density function of the infectious period distribution.

1. Non-Markov epidemic model

- The augmented likelihood is

$\pi(\mathbf{i}, \mathbf{r} \mid \beta, \theta, i_a, a) =$

$\qquad \prod_{j \neq a} \beta N^{-1} I(i_j-) \times \exp(- \beta N^{-1} \int S(t)I(t) \, dt)$

$\qquad\qquad\qquad \times \prod_{1 \leq j \leq n} f(r_j - i_j \mid \theta)$

where as before I(t-) means I(t) <u>just before</u> time t

1. Non-Markov epidemic model

- The target posterior density is

  $\pi(\beta, \theta, \mathbf{i}, i_a, a \mid \mathbf{r})$

  $\propto \pi(\mathbf{i}, \mathbf{r} \mid \beta, \theta, i_a, a) \, \pi(\beta, \theta, i_a, a)$

- Set independent priors as

  $\beta \sim$ Gamma(m_β, λ_β)

  $a \sim$ uniform on $\{1, \ldots, n\}$

  $i_a \sim$ uniform on $(-\infty, r_1)$

  $\theta \sim$?? Depends what θ is!

1. Non-Markov epidemic model

- MCMC algorithm that targets the posterior density needs ways of updating the parameters β, θ, $\mathbf{i}$, i_a and a.

- β, $\mathbf{i}$, i_a and a can all be updated as for the Markov model: β still has a Gamma-distributed full conditional distribution, and we can update infection times as before.

- Updates for θ depend on what θ is.

Contents

- 1. Non-Markov SIR epidemic model
- 2. <u>Debugging tips</u>
- 3. What to do with MCMC output
- 4. What can be estimated?

2. Debugging tips

1. Test each piece of code separately

Most MCMC algorithms in this field involve various components, e.g.

- Gibbs updates
- Metropolis-Hastings updates
- Likelihood

It is good practice to check each component works before proceeding.

2. Debugging tips

2. Validate output using simulations

As discussed in Lecture 1, one way to test MCMC code (e.g. for SIR model) is

- Simulate SIR model M times (e.g. M=1000)

- Run MCMC on each output to infer parameters

- Average parameter estimates from MCMC should be close to the known true values

2. Debugging tips

2. Validate output using simulations

If the MCMC code is time-consuming to run then an alternative is use simulation output that gives a single large epidemic - idea being that this should give reasonable information about the model parameters.

2. Debugging tips

3. Beware Zeroes

Some languages allow operations such as "0/0" without reporting an error.

2. Debugging tips

4. Try a very small data set

Sometimes it is possible to test MCMC code by using a very small data set where one can work out the required inference by hand. This can then be checked against the MCMC output.

2. Debugging tips

5. Use log likelihood

Many likelihoods require calculation of products which can in turn lead to numerical instabilities and run-time errors.

One way to tackle this issue is to instead work with the log likelihood, since

$$\log(A_1 \times A_2 \times ... \times A_m) = \log(A_1)+...+ \log(A_m)$$

2. Debugging tips

5. Use log likelihood (cont)

The likelihood may involve the calculation of Beta or Gamma functions.

R has built-in functions to compute such functions, i.e. beta, gamma; but if we are working on the log scale, instead of doing something like $\log(\text{gamma}(k))$ we could use another built in function $\text{lgamma}(k)$ to ensure numerical stability, especially if k is large.

Contents

- 1. Non-Markov SIR epidemic model
- 2. Debugging tips
- 3. <u>What to do with MCMC output</u>
- 4. What can be estimated?

3. What to do with MCMC output

In this section, for illustration it is assumed that we have MCMC output from the Markov SIR model removal-data-observed scenario:

$$(\beta_1, \gamma_1), (\beta_2, \gamma_2), \ldots, (\beta_M, \gamma_M)$$

where M is large (e.g. $M = 10^6$).

Each pair (β_k, γ_k) is (approx) a sample from the joint posterior density $\pi(\beta, \gamma \,|\, \mathbf{r})$

3. What to do with MCMC output

<u>Marginal summaries</u>

Quantities such as the marginal mean, median, variance etc of β and of γ can be readily obtained using the package R.

It is also useful to plot the marginal posterior density of each parameter and/or a histogram.

3. What to do with MCMC output

Joint summaries

It can be useful to assess the extent to which β and γ can be estimated separately.

The posterior correlation and a scatterplot of the samples against axes β and γ provide such information.

3. What to do with MCMC output

Functions of model parameters

The quantity R_0 is of enormous interest in mathematical epidemic theory. It is (roughly) defined as the average number of secondary cases caused by a typical infective in an infinitely large population of susceptibles.

If $R_0 \leq 1$, epidemics are unlikely to take off.

3. What to do with MCMC output

Functions of model parameters

For the (general) SIR model,

$$R_0 = \beta \, E(T_I),$$

where $E(T_I)$ is the mean infectious period.

This follows from the fact that each infective causes new infections at (Poisson) rate β during a period of time T_I .

3. What to do with MCMC output

<u>Functions of model parameters</u>

For the Markov model we have

$$R_0 = \beta \, E(T_I) = \beta \, / \, \gamma \, ,$$

since $T_I \sim Exp(\gamma)$.

Thus given the MCMC output we can create a new file containing

$(\beta_1/\gamma_1), (\beta_2/\gamma_2), \ldots, (\beta_M/\gamma_M)$

i.e. samples from the posterior density of R_0.

3. What to do with MCMC output

Functions of model parameters

R_0 can be summarised in the usual ways (mean, variance etc): also interesting to find the posterior probability that $R_0 \leq 1$.

3. What to do with MCMC output

Functions of model parameters

Can also be interesting to translate inference for rates into inference for probabilities.

e.g. $1 - \exp(-\beta / N)$ is the probability that one infective individual infects a given susceptible in one time unit.

Contents

4. What can be estimated?

As with any attempt to fit a model to data, it is always important to think about how informative the data are about the model parameters which one is trying to estimate.

In some settings it is obvious what can or cannot be estimated; in other settings it can be much less obvious.

4. What can be estimated?

Example: Latent periods

Consider the Markov SIR model with latent periods, say of fixed unknown length = c.
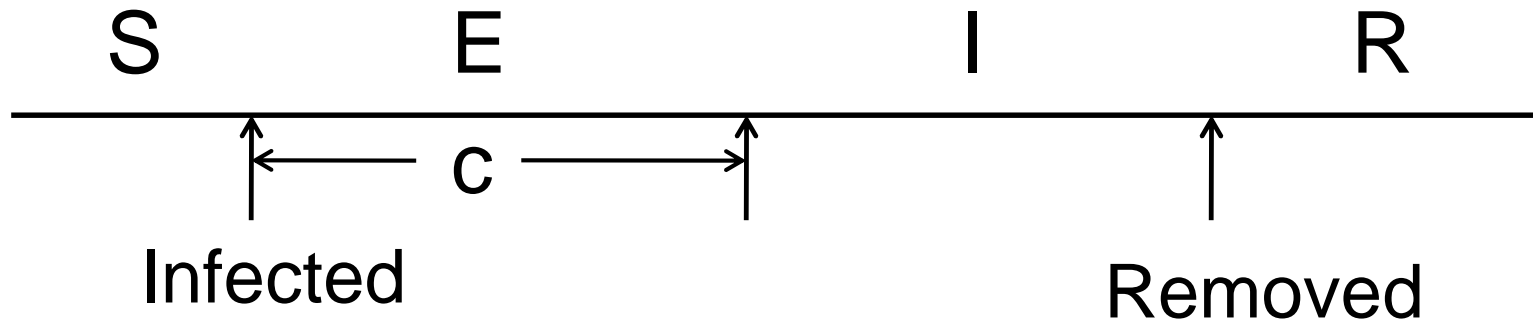
Thus when an individual is <u>infected</u>, they must wait c days until they become <u>infectious</u>.

Latent individuals are called "exposed".

4. What can be estimated?

Example: Latent periods

Thus we have an SEIR model where Exposed period = c time units.

4. What can be estimated?

Example: Latent periods

Introduce "exposure" times (= infection times)

$$e_1, e_2, \ldots, e_n, \text{ where } e_k = i_k - c$$

and define $\mathbf{e} = (e_1, e_{a-1}, e_{a+1}, \ldots, e_n)$.

As before, $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ is observed.

Define $\mathbf{i} = (i_1, i_2, \ldots, i_n)$.

4. What can be estimated?

The augmented likelihood is

$\pi(\mathbf{e, i, r} \mid \beta, \gamma, e_a, a) =$

$$\prod_{j \neq a} \beta N^{-1} I(i_j \text{-}) \times \exp(- \beta N^{-1} \int S(t)I(t) \, dt)$$
$$\times \gamma^n \exp\{- \gamma \sum (r_j - i_j)\}$$
$$\times 1_{\{e_k - i_k = c, \, k=1,\dots n\}}$$

4. What can be estimated?

Example: Latent periods

It is straightforward to adapt the standard MCMC algorithm to include c as an extra parameter - e.g. using M-H updates for c.

However such an algorithm would be uninformative about c given removal data alone.

4. What can be estimated?

Example: Latent periods

Roughly speaking, for any value of c, the infection rate β would be estimated accordingly - large c means large β and small c means small β .

So although the MCMC algorithm is correct, the output would need to be carefully interpreted. Here we would see high posterior correlation between c and β.

4. What can be estimated?

Example: Latent periods

In practice, a better strategy would be to fix c to certain (biologically reasonable) values and then perform estimation for β and γ.

4. What can be estimated?

Example: Gamma infectious periods

A common generalisation of the Markov SIR model is to have Gamma-distributed infectious periods.

Thus each infective remains so for a period of time T_I, where

$$T_I \sim \Gamma(c,d), \text{ say } (c = \text{shape}, d = \text{rate}).$$

Note $E(T_I) = c / d$.

4. What can be estimated?

Example: Gamma infectious periods

As seen earlier, the likelihood is

$\pi(\mathbf{i, r} \mid \beta, c, d, i_a, a) =$

$\qquad \prod_{j\neq a} \beta N^{-1} \, I(i_j\text{-}) \times exp(\text{-}\beta N^{-1} \int S(t)I(t)\, dt)$

$\qquad\qquad\qquad \times \prod_{1\leq j \leq n} f(r_j \text{-} i_j \mid c,d)$

where $\quad f(x \mid c,d) = x^{c\text{-}1}d^c \, exp(\text{-}dx) / \Gamma(c)$

is the p.d.f. of T_I .

4. What can be estimated?

Example: Gamma infectious periods

The two parameters c,d can be updated in an MCMC algorithm.

It is not immediately obvious if it is possible to estimate both parameters separately from removal data.

One might expect $E(T_I) = c / d$ to be estimated with reasonable precision.

4. What can be estimated?

Example: Gamma infectious periods

This suggests it might be better to use a different parameterisation of the Gamma distribution, with mean and variance rather than shape and rate.

4. What can be estimated?

Example: Data for Markov SIR model

Another important aspect of estimation is the detail of the data.

For example, suppose we have observations (= removal times) in a population of N=100 susceptibles, of whom n become infected.

Clearly if n=0, no inference can be drawn.

But what if n=1? n=10? n=100?